LAMP-TR-016 June 1998
UMIACS-TR-98-35
CS-TR-3911

# Enhancing Automatic Acquisition of Thematic Structure in a Large-Scale Lexicon for Mandarian Chinese

Mari Broman Olsen, Bonnie Dorr, Scott Thomas

Language and Media Processing Labratory
Instititue for Advanced Computer Studies
College Park, MD 20742

## Abstract

This paper describes a refinement to our procedure for porting lexical conceptual structure into new languages. Specifically we describe a two-step process for creating candidate thematic grids for Mandarin Chinese verbs, using the English verb heading the VP in the subdefinitions to separate senses, and roughly parsing the verb complement structure to match to our thematic structure templates. The procedure is part of a larger process of creating a usable lexicon for interlingual machine translation from a large on-line resource with both too much and too little information necessary for our system.

***The support of the LAMP Technical Report Series and the partial support of this research by the National Science Foundation under grant EIA0130422 and the Department of Defense under contract MDA9049-C6-1250 is gratefully acknowledged.

# Report Documentation Page

*Form Approved*
*OMB No. 0704-0188*



| 1. REPORT DATE | 2. REPORT TYPE | 3. DATES COVERED |
|---|---|---|
| **JUN 1998** | | **00-06-1998 to 00-06-1998** |

| 4. TITLE AND SUBTITLE | 5a. CONTRACT NUMBER |
|---|---|
| **Enhancing Automatic Acquisition of Thematic Structure in a Large-Scale Lexicon for Mandarian Chinese** | 5b. GRANT NUMBER |
| | 5c. PROGRAM ELEMENT NUMBER |
| 6. AUTHOR(S) | 5d. PROJECT NUMBER |
| | 5e. TASK NUMBER |
| | 5f. WORK UNIT NUMBER |
| 7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Language and Media Processing Laboratory,Institute for Advanced Computer Studies,University of Maryland,College Park,MD,20742-3275** | 8. PERFORMING ORGANIZATION REPORT NUMBER |
| 9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) | 10. SPONSOR/MONITOR'S ACRONYM(S) |
| | 11. SPONSOR/MONITOR'S REPORT NUMBER(S) |

12. DISTRIBUTION/AVAILABILITY STATEMENT
**Approved for public release; distribution unlimited**

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT **unclassified** | b. ABSTRACT **unclassified** | c. THIS PAGE **unclassified** | | **11** | |

# Enhancing Automatic Acquisition of Thematic Structure in a Large-Scale Lexicon for Mandarin Chinese

Mari Broman Olsen, Bonnie J. Dorr, and Scott C. Thomas

University of Maryland, College Park MD 20742, USA,
molsen,dorr,scthmas@umiacs.umd.edu
WWW home page: http://umiacs.umd.edu/labs/CLIP

**Abstract.** This paper describes a refinement to our procedure for porting lexical conceptual structure (LCS) into new languages. Specifically we describe a two-step process for creating candidate thematic grids for Mandarin Chinese verbs, using the English verb heading the VP in the subdefinitions to separate senses, and roughly parsing the verb complement structure to match thematic structure templates. We accomplished a substantial reduction in manual effort, without substantive loss. The procedure is part of a larger process of creating a usable lexicon for interlingual machine translation from a large on-line resource with both too much and too little information.

## 1 Introduction

In previous work on Spanish and Arabic (Dorr et al., 1997; Dorr, 1997a), we reported the results of an acquisition process for verb databases in new languages, using automatic assignment of candidate thematic structure templates ("grids") and manual verification of the output. This paper reports on acquisition of a Mandarin Chinese verb database from an on-line resource ten times as large as those used for Spanish and Arabic (600k, rather than 60k entries). The procedure is part of a larger process of creating a usable lexicon for interlingual machine translation from a large on-line resource with both too much and too little information necessary for our interlingual machine translation system (Dorr, 1997b; Hogan and Levin, 1994).

The major contributions of this work are: (i) reducing the effect of polysemy by addressing it in the preprocessing phase, and (ii) removing a substantial subset of automatically generated thematic grids requiring manual correction, by relating thematic information incorporated in Chinese verbs to overt complements in English. Both of the above result in a reduction of 11% of material that needs to be corrected by hand: 15,565 possible candidate thematic grids. Furthermore, the separation of senses allows candidate grids to be evaluated with respect to a particular sense. Noise that is introduced from polysemy in the English glosses—by putting the grids from the 'run a machine' and 'run a race' senses into a "bag of grids"—may therefore be eliminated, as grids are tied to

a particular sense. The reduction of manual effort has been accomplished without substantive loss of relevant definitions, as evaluated in a preliminary task, assigning thematic grids to verbs in 10 sentences from a corpus of 10 *Xinhua* articles.

We see this work as a step beyond that suggested by (Dorr et al., 1997), in which manual correction took place without first reducing the degree of ambiguity in the entry. Dorr et al. generated 18353 candidate thematic grids, representing 3623 verbs in the initial Spanish-English lexicon. Of that, 3025 entries were verified as correct (16.5%), and 15328 (83.5%) had to be modified in some way. There were 6082 deletions of entries, 334 reclassifications (resulting in changes of entries) and 6295 refinements of entries. The refinements included 3648 deletions of non-applicable entries; 2747 changes to prepositions, optional roles made obligatory, etc.; 2617 entries (955 verbs) deleted due to rarity of usage and/or disjointness with respect to WordNet concepts; 1213 new entries added (representing 1092 verbs not in the initial database). That is, there were a total of 9730 deletions, representing 63.5% of the required modifications, and 53% of the total number of candidate grids. Thus, an automatic process that reduces the number of deletions in a principled way would substantially reduce the manual correction process.

In the next section we describe the role of thematic grids in our system. We then describe our lexicon acquisition procedure, with respect to the verbs, detailing how we attempted to deal with polysemy and overgeneration of grids. We also report on other issues that arose in adapting the on-line resources.

## 2 Thematic Structure: Grids

Thematic structure serves as the interface between the syntactic component (parsing) and the lexical-semantic component, the Lexical Conceptual Structure (LCS). Verbs are assigned to classes that share syntactic and semantic behavior. That is, verbs in a class appear in the same types of sentences, with the same syntactic and semantic type of complements, represented as thematic (or "theta") roles. The syntactic and semantic behavior is abbreviated in the form of thematic grids, consisting of lists of obligatory and optional thematic roles, including agent, theme (patient), experiencer, source, goal and location.

In thematic grids, roles preceded by an underscore (_) are obligatory, and those preceded by a comma (,) are optional. A set of parentheses () indicates that the role must be expressed as a complement of a preposition or complementizer (e.g. the infinitival *to* in English). If the preposition is indicated, that preposition must be the head of the phrase. For example, the thematic grid `ag_th,src(from),goal(to)` indicates that agent and theme are obligatory, and source and goal are optional, and must be expressed by *from* and *to* prepositional phrases, respectively. Assigning this grid to the *Send* verbs, for example (class 11.1 in (Levin, 1993)), allows these verbs to appear in sentences like (1)-(4), but not (5)-(8), since the obligatory theme argument is missing.

(1) I sent the book.
(2) I sent the book to Mary
(3) I sent the book from the warehouse.
(4) I sent the book from the warehouse to Mary.
(5) * I sent.
(6) * I sent to Mary.
(7) * I sent from the warehouse.
(8) * I sent from the warehouse to Mary.

The thematic roles map directly into numbers, representing variables in the LCS. Although theta roles are theoretically unordered (Rappaport and Levin, 1988), the numbers correspond to a "canonical" linear position in a sentence and relative structural height in syntax and LCS trees. Thus `1` in the LCS corresponds to the **ag**(ent) thematic role and `2` to **th**(eme), since agents usually precede themes and occur higher in the syntactic tree. That is, in a sentence with an agent and a theme, typically the agent will be the subject and the theme the object, and both will precede other arguments.

The LCS for the above grid (simplifying irrelevant details) is given below: agent = **thing 1**, theme = **thing 2**, source preposition = **thing 3**, source complement of the preposition = **thing 4**, goal preposition = **thing 5**, goal complement of the preposition = **thing 6**. The * markers indicate where arguments are instantiated.

```
(cause (* thing 1)
  (go loc (* thing 2)
      ((* to 5) loc (thing 2) (at loc (thing 2) (thing 6)))
      ((* from 3) loc (thing 2) (at loc (thing 2) (thing 4))))
  (!!-ingly 26))
```

Thematic grids represent multiple structures; additionally, verbs in a language can take more than one thematic grid. For example, verbs like *fill, carpet, cloak* and *plug* allow the following grids, among others:[1]

(9) **_ag_th,mod-poss(with)** Derek filled the bucket with water.
(10) **_mod-poss_th** The water filled the bucket.

Other verb classes may take some of these grids but not others. Verbs like *inscribe, mark, sign, stamp* take the former, but not the latter, for example: *She signed his yearbook with her name*, but not **His name signed her yearbook*. The grids therefore group verbs by "semantic structure" (Levin and Rappaport Hovav, 1995). In contrast to "semantic content"—the idiosyncratic aspect of verb meaning—semantic structure determines syntactic patterning within and across languages (Dorr and Oard, 1998; Dorr and Katsova, 1998; Grimshaw, 1993; Pinker, 1984; Pinker, 1989).

[1] The **mod-poss** indicates a "possessed" item, paraphraseable by *have*: *The hole has water (in it).*

Most importantly for our system, the assignment of a set of thematic grids to a verb class allows us to create the interlingual LCS structures automatically (Dorr et al., 1995). Furthermore, in selecting grids for creating LCS entries for a new language, we leverage the fact that semantic structure overlaps across languages to a large degree. The task of creating the grids is therefore reduced to automatic generation, as described in Section 4, followed by manual correction to eliminate inappropriate grids (along with other modifications, described in Section 1 above). Before we describe the automatic process, we first describe some of the pre-processing required to extract appropriate verbs.

## 3 Verb Selection

The assignment of thematic grids/LCS structures to verbs is one step in the creation of a lexicon from a large (600k entries) machine readable Chinese-English dictionary. The dictionary was compiled by hand, by the Chinese-English Translation Assistance (CETA) group from some 250 dictionaries, some general purpose, others domain-specific or bilingual (Russian-Chinese, English Chinese, etc.). The CETA group, started in 1965 and continuing into the present decade, was a joint government-academic project. The machine-readable version of the CETA dictionary, *Optilex*, licensed from the MRM corporation, Kensington, MD.

CETA contains some information extraneous to our purposes. Some of the 250 resources used to create the dictionary were very domain-specific, including, for example, *Collier's North China Colloquial Collection*, a publication listing many regionalisms not observed anywhere else in China, and the *Faxue Cidian*, a dictionary of legal terms from Shanghai. We eliminated many archaic and technical verbs by eliminating verbs identified by CETA as derived from these sources.[2]

Even after archaic or idiosyncratic sources were eliminated, entries varied widely in specificity, from the general verbs (and other words) to the extremely specific, as the examples below show, given with the Pinyin, definition, and simplified character representation from CETA.

[2] BF Chinese-English Dictionary 1978, BE same as E but Chinese-Chinese 1978, AR Atlas of the PRC 1977 (for Chinese placenames), AO Gazetteer of the PRC (also for Chinese placenames), BQ extra new entries from the first two above BE and BF CJ standardized FBIS translations of Chinese communist terms, CM specialized terms extracted from Mao's works, CU Hong Kong glossary of Chinese communist terms, EJ 1981 idiom dictionary, EK 1982 idiom dictionary, FA Foreign Exchange terms 1963, IP International political economics glossary 1980, IQ Beijing social sciences academy economics terms 1983, NA world place names 1981, PP primary political economics glossary 1956, TM McGraw-Hill general scientific and technical dictionary 1963, VF Lin Yutang's dictionary 1972, VT 1973 Beijing foreign exchange glossary, WB Liang Shih-ch'iu's traditional dictionary 1973, YG Stanford's dictionary of Chinese communist terms 1973.

(11) po4_shi3 compel 迫使
(12) po4_shi3 force 迫使
(13) ben1_pao3 run 奔跑
(14) zou3 walk 走
(15) chu1_kou3 speak[3] 出口
(16) bi1_gong1 force_the_sovereign_to_abdicate 逼宫
(17) ben1_zou3_xiang1_gao4 run_around_spreading_the_news 奔走相告
(18) ca1_la5_zhe5_zou3 walk_dragging_one's_feet 擦拉着走
(19) chui1_xu1 speak_in_favor_of_somebody_in_exaggerated_terms 吹嘘

Although CETA is large, and in some ways exhaustive, some information required by our machine-translation lexicon is not directly encoded, notably part of speech.[4] We identified the verbs by a simple process. We parsed the DEF (gloss) field in the CETA entries from the selected sources. If the English glosses began with the infinitival 'to', the whole entry was used to generate as many new verb entries as there are verbs in the DEF field. As an example, the excerpt from following entry has four subentries in its DEF field. PY gives the Pinyin representation.[5]

PY: bian1 ta4
DEF: 1. to whip, to flog 2. <fig> to chastise, to castigate

After processing, each definition has a single subsense entry, i.e. there are four subentries.

## 4 Pairing Verbs and Thematic Grids

### 4.1 English glosses

For the Arabic and Spanish lexicon, we created candidate thematic grids by pairing target language words with the thematic grids associated with their English gloss, with manual correction over a period of two weeks. We did the same initial step for Chinese, as well. However, as described above, the senses had already been separated into different subentries. We thus had candidate thematic grid sets for each sense of a given verb.

The file containing Chinese grids was created by first matching the main verb of the English glosses to one or more entries in the English grids file. Separating

[3] As in 'to speak ill of someone.' This meaning is the first listed, although it is less common than others, including 'exit', as in exit signs (John Kovarik, p.c.).

[4] In fact, part of speech was not included in Chinese dictionaries at all, until the mid-80s (John Kovarik, Jin Tong p.c.); how and whether to do it is still controversial (`http://linguistlist.org/issues/9/9-1186.html`).

[5] CETA includes other fields not listed, including HWT and HWS encoding traditional and simplified characters, STC for the Standard Telegram Code, and REF for the dictionaries the entry came from.

polysemous entries is an aid to this process, since not all grids are associated with all verb senses. For example, a wide range of grids is available for the *Run* verbs. The first numbers, again, are classes from Levin (Levin, 1993). Numbers less than 9 are classes not found in Levin that were created automatically (Dorr, 1997b).

(20) 26.3 `_ag` 持 chi2 run
(21) 26.3 `_ag_ben_th` 持 chi2 run
(22) 26.3 `_ag_th,ben(for)` 持 chi2 run
(23) 47.5.1 `_ag,mod-loc()` 持 chi2 run
(24) 47.5.1 `_loc_th` 持 chi2 run
(25) 47.5.1 `_th_loc()` 持 chi2 run
(26) 47.7 `_th_goal()` 持 chi2 run
(27) 47.7 `_th_src(from)_goal(to)` 持 chi2 run
(28) 51.3.2 `_ag` 持 chi2 run
(29) 51.3.2 `_th,src(),goal()` 持 chi2 run

In contrast, a relatively small number is available for other meanings of this character.

(30) 31.2 `_exp_perc,mod-poss(in)` 持 chi2 support
(31) 47.8 `_th_loc` 持 chi2 support

In previous work, all grids were associated with a single entry and the checker was presented with a "bag of grids", without a link to a specific meaning. Since manual separation of senses was necessary, the likelihood of human error was high: checkers would delete or retain grids depending on which sense of the verb they had in mind. In the case at hand, it turns out that 持 chi2 means 'run', as in 'run a business' or 'run a machine', whereas the theta grids were derived from the motion verb *run* in English. Should the grids prove inappropriate in the manual verification stage, they can be deleted without affecting entries with other meanings.

### 4.2 Automatic Modification of Candidate Grids

Each thematic grid in the initial candidate set describes the argument structure for the head verb of the gloss, in some usage of that (English) verb. To construct appropriate LCSs for the Chinese verb, these grids must be manually checked and modified where necessary. We have further parsed the DEF field to automatically make certain modifications that in earlier work had been done by hand.

For instance, the candidate set for the Chinese verb in (16) above, glossed 'to force the sovereign to abdicate,' contains the grid `_ag_th,prop(to)`, because the English verb *force* takes an agent, theme and optional propositional complement. After parsing the gloss into subphrases, we can posit that 'the sovereign' is theme, and 'to abdicate', the propositional element. On the assumption that a gloss of

this sort implies that theme and propositional element are part of the Chinese verb meaning and *not* expressed as overt complements, the grid is reduced to `_ag`; 'the sovereign' and 'to abdicate' are set aside, to be inserted directly into the LCS for the Chinese entry. Thus, for hand checking, we construct a grid that appears like:

(32) 002 `_ag` 逼宫 bi1_gong1 force_the_sovereign_to_abdicate
(th = sovereign) (prop = to_abdicate)

Similarly, the following Chinese word receives the grid shown in (9), but with the possessional modifier `mod-poss(with)` lexicalized by the verb itself, and thus removed from the grid:

(33) 9.8 `_ag_th` 填土 tian2_tu3 fill_in_with_earth (mod-poss = earth)

The underlying intuition is that verbs that incorporate thematic elements in their meaning would not allow that element to appear in the complement structure: **fill_in_with_earth with gravel*, cf. English **I sodded my lawn with ivy*[6].

The matching of gloss verb-complements to thematic roles is made as follows. We first parsed the gloss with simple context-free phrase-structure rules, to retrieve a flat structure, consisting of the V and a list of complements: NPs, PPs, clauses like 'to abdicate', and predicate adverbs or adjectives, like 'weary' in 'to be weary'. PPs headed with 'of' were attached low and not considered as a VP complement, e.g. *give* [*an explanation of the situation*].

Information in parentheses was ignored. Thus, had the gloss above been 'to force (i.e. the sovereign) to abdicate', we would have assumed that the Chinese verb *required* a theme argument (like 'the sovereign'), and the grid would have been `_ag_th` instead of just `_ag`. Parentheses do contain some apparently important material. For example, there is a gloss 'to kill (or catch) a tiger', which appears to condense two different senses. However, this usage of parentheses was mostly found in the sections of CETA we suppressed. A series of nouns was considered a single NP, as in 案检: an4 jian3; DEF: to investigate [a law case].

Having split the gloss into its thematic parts, we then match the PPs to thematic roles that specify the same head as the PP, and match propositional elements that have matching prepositions or particles (i.e. the 'to' in 'to abdicate'). Some grids specify roles with no particular preposition, in which case we heuristically assign roles according to this table:

from: `src` (source) or `instr`
for: `purp` (purpose)
with: `instr` or `mod-poss`
without: `mod-poss`
into, to against: `goal`
under, around, along: `mod-loc`

[6] We are ignoring 'cognate objects', as in *I sodded my lawn with the best sod available* (Macfarland, 1995)

*Adverbs* become **manner** components, in positions where they typically modify the verb ('to blindly worship foreign things'), rather than an adjective ('to be seriously ill'). The adverbial manner components become part of the LCS, if the entry passes through the hand inspection phase. A gloss that ends with a *dangling preposition* is taken as a sign that, where the English verb takes a PP, the Chinese verb fills the same role with a bare NP argument. Thus the parentheses are removed from the grid for that role (see Section 2). *Bare noun phrases* are matched to non-prepositional-phrase thematic roles. *Predicate adjectives* match **pred**, an identificational predicate—in this case, naming a property. Any material in the gloss not matching anything in the thematic grid is kept for incorporation into the LCS as a modifier.

In this manner, 11360 distinct theta role assignments were created. In some cases the original theta-roles list actually becomes empty, in which case it appears as **_0**, the thematic grid for verbs with no semantic arguments, such as *rain* in English *It's raining*.

After we saturate the relevant components of the thematic grids, we use the filled grids to reduce the candidate set of grids. If the set of theta roles lexicalized by a Chinese verb sense for one candidate grid (which may be the empty set) is a proper subset of that for another grid of that verb sense, then the smaller grid is discarded, resulting in an a 11% reduction in the number of entries that need to be hand-checked. Thus, if there were a thematic grid **_ag_th** generated for 'to force the sovereign to abdicate', it would be discarded in favor of the grid above.

Similarly, the seven candidate grids for 'to serve as a guide' reduce to one, since only one could incorporate the predicate with 'as', that from Levin class 29.6, which includes verbs like *act, behave*, and *pose* as well as *serve*:

(34) Entry HWS: 充向导
PY: chong1 xiang4 dao3
DEF: 'to serve as a guide'
(35) Retained:
29.6.b **_th_pred(as)**
(36) Suppressed:
13.1 **_ag_goal_th** (e.g. *We served them food*)
13.1 **_ag_th_goal(to)** (e.g. *We served food to them*)
13.4.1 **_ag_th,mod-poss(with)** (e.g. *I served him with a warrant*)
13.4.1 **_ag_th,goal(to)** (e.g. *We served a warrant to them*)
54.3 **_ag_th_loc()** (e.g. *We serve 114 people in this restaurant*)
54.3 **_th_poss** (e.g. *This restaurant serves 114 people*)

## 5 Results

Using the process described above, 15565 thematic grids were eliminated, representing 11% of the total number of candidates. We began the process of manual evaluation of the theta grids, beginning with the verbs in 10 articles from *Xinhua*,

comparable to the *Wall Street Journal* in content: 124 grids were suppressed for 47 verbs (29 classes), leaving 3041 grids for 263 verbs (characters, rather than definitions). A set of 51 theta grids were generated for the 13 verbs in ten sentences from these articles. Chinese speakers deleted 17 grids, or 33.3%. Although these results are a tiny subset of the full verb lexicon, this figure compares favorably to the 53% deletion required of the Spanish data. Importantly, none of the relevant grids had been discarded by our algorithm.[7]

The fact that we parsed the complement structure in the subentries alerted the language experts (John Kovarik and Mary-Ellen Okurowski, both from the Department of Defense (DOD), and Ron Dolan from the Library of Congress) to additional senses that should be eliminated from the lexicon, e.g. those not properly considered verbal. Furthermore, although we used only the most general sources, all the dictionaries included entries from both classical and colloquial Chinese. Only the latter is used in our domain. Additionally, the classical entries are often archaisms and figurative uses that would likely not have the same thematic structure, for example, the meaning 'to shelve' derived from a character meaning 'to push (down)' (PY an4). In addition, the syntactic structure assigned to the gloss demonstrated that some of the entries glossed as verbs are more appropriately treated as prepositions or prepositional phrases. Removing old and syntactically incorrect entries resulted in a further 40% reduction for verb senses in the ten articles. Since we have been generating an average of 3.3 thematic grids per sense, we have decided to do preprocessing before generating the other candidate grid sets.

## 6 Conclusions and Future Work

We have described a procedure for automatically reducing the amount of manual checking necessary for building the thematic grid structure for verbs in Chinese. We anticipate that this procedure will save us time over our original checking procedure. The latter, in turn, reduced the amount of time required to create thematic structure from 6 person months (for a lexicon with 60k entries and 3-4k verbs) to approximately two weeks of hand verification. The time savings for our project is even more imperative, since we expect to have almost double that size in verbs alone, even after removing inappropriate entries. The procedure described in this paper provides further streamlining for the process of acquiring large-scale lexica for NLP applications with non-optimal on-line resources.

In addressing the polysemy problem in this context, we have, as a by-product, produced a sense-to-syntax mapping, tying a verb sense/character pair to a set of grids representing syntactic as well as semantic structure. This mapping, in turn, could be used not only for machine translation, but for segmentation and word sense disambiguation algorithms for Chinese.

[7] The copular grid for the verb *shi4* was added to the set, using a grid assigned to other copular verbs, namely *wei2* and *zuo4*. Somewhat surprisingly, the absence of the copular grid in our candidate set resulted from an absence in CETA of the copular meaning for that verb.

## Acknowledgments

This work has been supported, in part, by DOD Contract MDA904-96-C-1250. The second author is also supported by DARPA/ITO Contract N66001-97-C-8540, Army Research Laboratory contract DAAL01-97-C-0042, NSF PFF IRI-9629108 and Logos Corporation, NSF CNRS INT-9314583, and Alfred P. Sloan Research Fellowship Award BR3336. We would like to thank members of the following lab groups at Maryland: Computational Linguistics and Information Processing (CLIP), and Language And Media Processing (LAMP), particularly Galen Wilkerson for his implementation and description of verb selection, and John Kovarik, a Chinese language instructor on loan from the DOD.

## References

Dorr, B. J. (1997a). Large-Scale Acquisition of LCS-Based Lexicons for Foreign Language Tutoring. In *Proceedings of the ACL Fifth Conference on Applied Natural Language Processing (ANLP)*, pages 139–146, Washington, DC.

Dorr, B. J. (1997b). Large-Scale Dictionary Construction for Foreign Language Tutoring and Interlingual Machine Translation. *Machine Translation*, 12(4):271–322.

Dorr, B. J., Garman, J., and Weinberg, A. (1995). From Syntactic Encodings to Thematic Roles: Building Lexical Entries for Interlingual MT. *Machine Translation*, 9:71–100.

Dorr, B. J. and Katsova, M. (1998). Lexical Selection for Cross-Language Applications: Combining LCS with WordNet. In *Proceedings of AMTA-98*, Lanhorne, PA.

Dorr, B. J., Marti, A., and Castellon, I. (1997). Spanish EuroWordNet and LCS-Based Interlingual MT. In *Proceedings of the MT Summit Workshop on Interlinguas in MT*, San Diego, CA.

Dorr, B. J. and Oard, D. W. (1998). Evaluating resources for query translation in cross-language information retrieval. In *Proceedings of the First International Conference on Language Resources and Evaluation*, Granada, Spain.

Grimshaw, J. (1993). Semantic Structure and Semantic Content in Lexical Representation. unpublished ms., Rutgers University, New Brunswick, NJ.

Hogan, C. and Levin, L. (1994). Data Sparseness in the Acquisition of Syntax-Semantics Mappings. In *Proceedings of the Post-COLING94 International Workshop on Directions of Lexical Research*, pages 153–159, Nicoletta Calzolari and Chengming Guo (co-chairs), Tshinghua University, Beijing.

Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation.* University of Chicago Press, Chicago, IL.

Levin, B. and Rappaport Hovav, M., editors (1995). *Unaccusativity: At the Syntax-Lexical Semantics Interface.* The MIT Press, Cambridge, MA. LI Monograph 26.

Macfarland, T. (1995). *Cognate Objects and the Argument/Adjunct Distinction in English.* PhD thesis, Northwestern University, Evanston, IL.

Pinker, S. (1984). *Language Learnability and Language Development.* MIT Press, Cambridge, MA.

Pinker, S. (1989). *Learnability and Cognition: The Acquisition of Argument Structure.* The MIT Press, Cambridge, MA.

Rappaport, M. and Levin, B. (1988). What to do with $\theta$-Roles. In Wilkins, W., editor, *Syntax and Semantics: Vol. 21, Thematic Relations*, pages 7–36. Academic Press, New York.